IX. Monika Kolvenbach

DAS MORPHOLOGISCHE LEXIKON (MOLEX) DES SYSTEMS PLIDIS

bzw. "Wie man die morphologische Analyse in einem auf Dialog mit der Maschine eingestellten System durch reine Lexikonaufsuche ersetzt"

Während des Jahres 1970 konkretisierte sich die Idee, daß es sinnvoll sei, eine Möglichkeit zu haben, als "Normalverbraucher" ohne Programmierkenntnisse und DV-Erfahrung mit einem Informationssystem kommunizieren zu können, da beim damaligen Stand der Entwicklung die Benutzung eines Informationssystems nur über den Filter von DV-Fachleuten laufen konnte, die die Kenntnisse hatten, mit einem System in der von diesem System verlangten Sprache umzugehen.

Wollte man jedoch eine natürlichsprachliche Zugangsmöglichkeit zu Informationssystemen schaffen, dann war es unerläßlich, daß man einem solchen System die deutsche Sprache beibrachte. Dieses "Beibringen der deutschen Sprache" hatte in etwa so vor sich zu gehen, wie man einem Ausländer (der in diesem Fall noch extrem dumm und zumindest aus einem völlig entlegenden Zipfel der Erde kommen müßte, da er keinerlei Kenntnisse über unsere "deutschen" nur mitgedachten, aber nicht formulierten sprachlichen Zusammenhänge haben dürfte) die deutsche Sprache beibringt.

Um die deutsche Sprache zu verstehen, mußte man deshalb Komplexe einbauen, die das Verständnis der Morphologie, der Syntax und der Semantik vermitteln. Diese drei Komponenten sind als solche für den Verstehensprozeß unabdingbar; wie man jedoch diese im einzelnen realisiert, ist eine Frage, die in Bezug zu dem jeweiligen "Schüler" zu beantworten ist.

Da in unserem Fall der Schüler der Computer war, mußte das Vorgehen sich teilweise von dem Vorgehen bei einem menschlichen Schüler entfernen, da man bei diesem das Verständnis irgendeiner Sprache und somit auch sprachlicher Zusammenhänge voraussetzen

kann, der Computer jedoch bis dahin in einem natürlichsprachlichen Sinn "sprachlos" war. Aus dem Grund der "Sprachlosigkeit" im oben beschriebenen Sinn konnte man mit Sicherheit davon ausgehen, daß eine Vielzahl von "Verständnisschwierigkeiten" auf allen drei Ebenen auftauchen würde. Um diese Fehler leichter lokalisieren zu können, schien es sinnvoll, die drei notwendigen Ebenen sauber zu trennen, d.h., einen streng modularen Aufbau zu wählen.

Zentraler Punkt in diesem "Sprachverstehen" war die Syntaxanalyse, da es sicher sein mußte, daß bei dieser Unkenntnis des Schülers eine vollständige Vermittlung der Semantik einer Sprache nicht möglich sein konnte, sondern es sich immer nur um die Semantik eines streng umgrenzten Fachgebiets oder Weltausschnitts handeln konnte; denn eine Vermittlung der gesamten Semantik einer Sprache ist selbst bei spracherfahrenen Schülern schon äußerst schwierig.

Vorbedingung für die Syntaxanalyse ist jedoch eine morphologische Analyse, d.h. die Zuordnung von verschiedenen Wortformen zu ihren Grundformen. So mußte etwa die Wortform "GEHST" als eine spezielle Form des Infinitivs "GEHEN" erkannt werden, um über diese Wort-Grundform "GEHEN" einmal die Zugehörigkeit der Wortform "GEHST" zu der Klasse der Verben zu vollziehen, und zum anderen für den Weltausschnitt notwendige semantische Informationen an diese Grundform und nicht an jede mögliche Flexionsform anbinden zu können. Bei bezüglich ihrer Wortform mehrdeutigen natürlichsprachlichen Ausdrücken, wie etwa "SCHLOSS" - im vorliegenden Fall wurde von der zu distinktiven Zwecken heranziehbaren Großschreibung abgesehen - mußte die morphologische Analyse sowohl die mögliche Beziehung zu dem Verb "SCHLIESSEN" als auch zu dem - semantisch ebenfalls mehrdeutigen - "SCHLOSS" als Nomen aufzeigen.

Um dies zu erreichen, wäre es einmal möglich gewesen, aufgrund der vorliegenden Textwörter eine morphologische Analyse vorzunehmen und von daher zu den gewünschten Ergebnissen zu kommen. Ein solches Verfahren wurde in den Anfängen des Projekts gewählt und in den Forschungsberichten Nr. 18.1, 18.2 und 19 des Instituts für deutsche Sprache (Arbeitsgruppe MasA: Zur maschinellen Syntax-

analyse I und II; morphosyntaktische Voraussetzungen für eine maschinelle Sprachanalyse des Deutschen. Hrsg. U. Engel, I. Vogel, Mannheim 1974) beschrieben. Ebenfalls für dieses Verfahren entschied sich eine Arbeitsgruppe in Saarbrücken (dokumentiert in: "Elektronische Syntaxanalyse der deutschen Gegenwartssprache" von Hans Eggers, Max Niemeyer Verlag Tübingen, 1969).

Da die im MasA-Bericht beschriebene morphologische Analyse sehr zeitaufwendig ist, entschied man sich innerhalb der Abteilung LDV im nachhinein für ein anderes Verfahren. Man konstruierte ein Lexikon - eben das besagte MOLEX - in dem sämtliche möglichen Ergebnisse einer solchen morphologischen Analyse für die dort aufgenommenen Wortformen der deutschen Sprache festgehalten wurden.

Hierbei wurde Wert darauf gelegt, daß alle Wortformen, die in dieses MOLEX aufgenommen wurden, morphologisch vollständig beschrieben waren, ganz gleich, ob der Anwendungsbereich diese Vollständigkeit verlangte oder nicht. Begründung hierfür war, daß man bei einem Übergang zu einem anderen Weltausschnitt wohl gezwungen sein würde, die semantischen Einträge zu den einzelnen Grundformen zu ändern, daß die morphologischen und syntaktischen Beschreibungen in diesem Fall jedoch Bestand haben sollten, wenn z.B. auch noch bestimmte syntaktische Konstruktionen als nicht projektrelevant aufgeklammert würden und auf das Abprüfen bestimmter Fehlerquellen verzichtet würde, da diese im Weltausschnitt nicht belegt waren.

Das MOLEX in seiner jetzigen Form soll zur Abwasserüberwachung von metallerarbeitenden Betrieben im Land Baden-Württemberg eingesetzt werden. Anwendungsgebiet für "AUS" in diesem Weltausschnitt ist deshalb höchstens seine Funktion als Präposition, in Sätzen wie: "Gibt es Proben aus Mannheim, in denen ein Grenzwert für Arsen verletzt wurde?". Trotzdem wurde "AUS" im MOLEX sowohl als Präposition - wie für den Anwendungsbereich notwendig - notiert, als auch als abtrennbares Verbpräfix für Sätze wie

"Er geht aus." als auch als Nomen wie für Sätze "Der Ball war im Aus.".

Da das Auffüllen des Lexikons von Hand sehr zeitaufwendig war und die morphologische Analyse innerhalb des Projekts arbeitsaufwandsmäßig keinen zentralen Platz, wie etwa die semantische Analyse, einnehmen sollte, wurde ein MOLEX-Generator entwickelt, der aus der Angabe einer Grundform und einer Wortklassen-Nummer sämtliche MOLEX-Einträge in der erforderten Form generiert. Grundlage hierfür war die Wortklasseneinteilung des Wahrig (Deutsches Wörterbuch, Gütersloh, Berlin, München, Wien, 1968, 1973), der zu den Wortklassen auch Flexionsklassen angibt. Diese Flexionsklassen wurden, soweit notwendig, erweitert und teilweise auch verändert. Der Wortbestand des Wahrig (s.o.) wurde auf Datenträger übernommen und mit den für die Generierung notwendigen - teilweise veränderten und ergänzten - Flexionsklassennummern versehen. Dadurch wird es dem IdS möglich sein, im Laufe des Jahres 1980 ein morphologisches Lexikon zur Verfügung zu stellen, das sowohl den Sprachumfang des Wahrig (s.o.) als auch den des Anwendungsgebiets des Projekts PLIDIS anbietet. Eine Erweiterung des Sprachumfangs dieses Lexikons - z.B. auf den Sprachumfang des Mannheimer Corpus der geschriebenen Sprache ist insofern problemlos, als es nur darum geht, nicht im MOLEX verzeichnete und durch das Stichwortregister leicht zu ermittelnde Wortformen in ihrer Grundform mit einer Flexionsklassennummer zu versehen und dann generieren und in das MOLEX einspielen zu lassen.

Dadurch wird es möglich, morphologisch - wenn auch in der vollen Vielfalt der Möglichkeiten - annotierte Texte des Mannheimer Corpus herzustellen.

Der Einsatz des Syntax-Parsers auf diesen Texten wird dann ebenfalls möglich, wenn bestimmte Routinen, die für die Texte des Mannheimer Corpus von Wichtigkeit sind, für den Anwendungsbereich jedoch ohne Belang waren - wie etwa Pronominalisierungen, Konsistenzprüfungen der unterschiedlichsten Art - (obwohl der Syntax-Parser allgemein ausgelegt ist, geht er bei der Person/Numerus-

Beschreibung z.B. immer davon aus, daß es sich, wenn möglich, um 3. Person Singular oder Plural handelt) - eingebaut sein werden.

Bis zu diesem Schritt, einer syntaktischen annotierten Belegsammlung aus dem Mannheimer Corpus der geschriebenen Sprache sind jedoch noch eine Vielzahl von Veränderungen vorzunehmen, die, da es sich dabei um komplexe Probleme - wie etwa die Konsistenzprüfung handelt - noch einen großen Arbeitsaufwand erfordern.

Wie der Molex-Generator arbeitet, soll das folgende Beispiel zeigen:

Das Entstehen eines MOLEX-Eintrags mithilfe
des Generierungsprogramms
oder: Wie man aus einer Zeile automatisch viele Zeilen macht

1. Eintrag in der Eingabedatei "F"

```
.....
(fahren (V 130))
.....
```

2. Paradigma für die Verbklasse 130

```
(130 1 A 3 FAHR (1 2 3 4 12 13 14 15 16 17 18 27 28))
(130 2 A 6 FAEHR (6 29))
(130 3 A 8 FUHR (43 44 45 46 47 48))
(130 4 A 9 FUEHR (49 50 51 52 53 54))
```

(Die Zahlen und Buchstaben zwischen Klassennummer und Stamm sind interne Prüfnummern, an denen z.B. nachprüfbar ist, ob die Generierungsnummern für die zu bildenden Tempora zugelassen sind, u.a.m.)

3. Generierungsnummern (insoweit als sie zum Verständnis von 2. notwendig sind)

```
(1 ENDUNG -E PN 1 TEMP GE MOD IND DIA AKT)
(2 ENDUNG -E PN (1 3) TEMP GE MOD KONJ DIA AKT)
(3 ENDUNG -EN PN (4 6) TEMP GE MOD (IND KONJ) DIA AKT)
(4 ENDUNG -EST PN 2 TEMP GE MOD KONJ DIA AKT)
(5 ENDUNG -ET PN 5 TEMP GE MOD IND DIA AKT)
(6 ENDUNG -ST PN 2 TEMP GE MOD IND DIA AKT)
(12 ENDUNG -END TEMP P1)
(13 ENDUNG -EN TEMP IN)
(14 ENDUNG -EN PN (2 5) TEMP GE MOD BEF DIA AKT)
(15 ENDUNG NIL PN 2 TEMP GE MOD BEF DIA AKT)
(16 ENDUNG -E PN 2 TEMP GE MOD BEF DIA AKT)
(17 ENDUNG -T PN 5 TEMP GE MOD BEF DIA AKT)
(18 ENDUNG -ET PN 5 TEMP GE MOD BEF DIA AKT)
(27 ENDUNG -T PN 5 TEMP GE MOD IND DIA AKT)
(28 ENDUNG -ET PN 5 TEMP GE MOD KONJ DIA AKT)
(29 ENDUNG -T PN 3 TEMP GE MOD IND DIA AKT)
(43 ENDUNG NIL PN (1 3) TEMP VE MOD IND DIA AKT)
(44 ENDUNG -ST PN 2 TEMP VE MOD IND DIA AKT)
(45 ENDUNG -EN PN (4 6) TEMP VE MOD IND DIA AKT)
(46 ENDUNG -T PN 5 TEMP VE MOD IND DIA AKT)
(47 ENDUNG -EST PN 2 TEMP VE MOD IND DIA AKT)
(48 ENDUNG -ET PN 5 TEMP VE MOD IND DIA AKT)
(49 ENDUNG -E PN (1 3) TEMP VE MOD KONJ DIA AKT)
(50 ENDUNG -EST PN 2 TEMP VE MOD KONJ DIA AKT)
(51 ENDUNG -ET PN 5 TEMP VE MOD KONJ DIA AKT)
(52 ENDUNG -EN PN (4 6) TEMP VE MOD KONJ DIA AKT)
(53 ENDUNG -ST PN 2 TEMP VE MOD KONJ DIA AKT)
(54 ENDUNG -T PN 5 TEMP VE MOD KONJ DIA AKT)
```

4. Stammänderungsliste, wie aus 2. ableitbar:

-ah- --➤ -aeh-; -ah- --➤ -uh-; -ah- --➤ -ueh-

5. Generierte MOLEX-Einträge

```
(FAHRE (VERB NF FAHREN KNG 7943 PN 1 TEMP GE MOD IND DIA AKT))
(FAHRT (VERB NF FAHREN KNG 8007 PN (1 3) TEMP GE MOD KONJ DIA AKT))
(FAHREN (VERB NF FAHREN KNG 7727 PN (4 6) TEMP GE MOD (IND KONJ) DIA AKT))
(FAHREST (VERB NF FAHREN KNG 7815 PN 2 TEMP GE MOD KONJ DIA AKT))
(FAHREND (VERB NF FAHREN KNG 8191 TEMP P1))
(FAHREN (VERB NF FAHREN KNG 8191 TEMP IN))
(FAHREN (VERB NF FAHREN KNG 7831 PN (2 5) TEMP GE MOD BEF DIA AKT))
(FAHR (VERB NF FAHREN KNG 7815 PN 2 TEMP GE MOD BEF DIA AKT))
(FAHRE (VERB NF FAHREN KNG 7815 PN 2 TEMP GE MOD BEF DIA AKT))
(FAHRT (VERB NF FAHREN KNG 7703 PN 5 TEMP GE MOD BEF DIA AKT))
(FAHRET (VERB NF FAHREN KNG 7703 PN 5 TEMP GE MOD BEF DIA AKT))
(FAHRT (VERB NF FAHREN KNG 7703 PN 5 TEMP GE MOD IND DIA AKT))
(FAHRET (VERB NF FAHREN KNG 7703 PN 5 TEMP GE MOD KONJ DIA AKT))
(FAEHRST (VERB NF FAHREN KNG 7815 PN 2 TEMP GE MOD IND DIA AKT))
(FAEHRT (VERB NF FAHREN KNG 7751 PN 3 TEMP GE MOD IND DIA AKT))
(FUHR (VERB NF FAHREN KNG 8007 PN (1 3) TEMP VE MOD IND DIA AKT))
(FUHRST (VERB NF FAHREN KNG 7815 PN 2 TEMP VE MOD IND DIA AKT))
(FUHREN (VERB NF FAHREN KNG 7727 PN (4 6) TEMP VE MOD IND DIA AKT))
(FUHRT (VERB NF FAHREN KNG 7703 PN 5 TEMP VE MOD IND DIA AKT))
(FUHREST (VERB NF FAHREN KNG 7815 PN 2 TEMP VE MOD IND DIA AKT))
(FUHRET (VERB NF FAHREN KNG 7703 PN 5 TEMP VE MOD IND DIA AKT))
(FUEHRE (VERB NF FAHREN KNG 8007 PN (1 3) TEMP VE MOD KONJ DIA AKT))
(FUEHREST (VERB NF FAHREN KNG 7815 PN 2 TEMP VE MOD KONJ DIA AKT))
(FUEHRET (VERB NF FAHREN KNG 7703 PN 5 TEMP VE MOD KONJ DIA AKT))
(FUEHREN (VERB NF FAHREN KNG 7727 PN (4 6) TEMP VE MOD KONJ DIA AKT))
(FUEHRST (VERB NF FAHREN KNG 7815 PN 2 TEMP VE MOD KONJ DIA AKT))
(FUEHRT (VERB NF FAHREN KNG 7703 PN 5 TEMP VE MOD KONJ DIA AKT))
```

Legende: NF = Normalform
KNG = Kasus, Numerus, Genus (eindeutig verschlüsselt)
PN = Person, Numerus
TEMP = TEMPus
GE = GEgenwart
VE = VErgangenheit
MOD = MODus
BEF = BEFehl
IND = INDikativ
KONJ = KONJunktiv
DIA = DIAthese
AKT = AKTiv

Über die Generierung des Verbs FAHREN werden somit noch die folgenden Lexeme in das MOLEX mit allen ihren Flexionsformen eingetragen: die Nomen FAHRT, FUHRE, FAHRENDE(R) und das Verb FÜHREN. Damit ist sichergestellt, daß alle morphologischen Beschreibungen zu den einzelnen Wortformen vorhanden sind. Diese Erweiterungen erfolgen jedoch nicht automatisch, sondern nur DV-unterstützt.